Nucleic Acids Research, 2004, Vol. 32, Database issue ***D251–D254***
DOI: 10.1093/nar/gkh060

TOPS: an enhanced database of protein structural topology

Ioannis Michalopoulos, Gilleain M. Torrance[1], David R. Gilbert[1] and David R. Westhead*

School of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK and [1]Department of Computing Science, University of Glasgow, 17 Lilybank Gardens, Glasgow G12 8QQ, UK

Received August 15, 2003; Revised and Accepted September 27, 2003

ABSTRACT

The TOPS database holds topological descriptions of protein structures. These compact and highly abstract descriptions reduce the protein fold to a sequence of Secondary Structure Elements (SSEs) and three sets of pairwise relationships between them, hydrogen bonds relating parallel and anti-parallel β strands, spatial adjacencies relating neighbouring SSEs, and the chiralities of selected supersecondary structures, including connections in βαβ units and between parallel α helices. The database is used as a resource for visualizing folding topologies, fast topological pattern searching and structure comparison. Here, significant enhancements to the TOPS database are described. The topological description has been enhanced to include packing relationships between helices, which significantly improves the description of protein folds with little β strand content. Further, the topological description has been annotated with sequence information. The query interfaces to the database have been improved and the new version can be found at http://www.tops.leeds.ac.uk/.

INTRODUCTION

TOPS is the collective name for a set of tools associated with topological descriptions of protein 3D folds. These include the TOPS program (1), which produces a visualization aid for folding topology known as a TOPS cartoon, and several tools that convert information from the TOPS program into a formal description of folding topology known as a TOPS diagram (2) (see Fig. 1 for examples). These tools were used to create the original TOPS database (3) (http://tops.ebi.ac.uk/tops), including the TOPS Atlas of high quality TOPS cartoons, searches for user-defined topological patterns over all structures from the Protein Data Bank (PDB) (4) and facilities for structure comparison at the topological level. These facilities have been used extensively by the structural biology community.

The TOPS diagram (Fig. 1) is a compact and highly abstract description that, with the original definition (2), reduces the protein fold to a sequence of secondary structure elements (SSEs) and three sets of pairwise relationships between SSEs, hydrogen bonds relating parallel and anti-parallel β strands, spatial adjacencies relating neighbouring SSEs, and the chiralities of selected supersecondary structures, including connections in βαβ units and between parallel α helices. This topological description is very effective for protein folds with substantial β sheet content, owing to the well-defined parallel and anti-parallel adjacency of β strands in β sheet structures, and the tendency of some α/β folds to form well-defined layers with helices packing above or below parallel β sheets according to the chirality of the supersecondary connection between parallel β strands. However, the description is much less useful for protein folds comprising mostly α helices, where the only information enabling the method to distinguish between different all α folds is a relatively sparse set of chirality relationships, pertaining to connections between near-parallel helices.

Here we describe new versions of the TOPS database and associated tools with significantly enhanced functionality. In order to improve the treatment of all α folds we have augmented the TOPS diagram to include helix–helix packing relationships and angles, a more extensive set of supersecondary structure chiralities, and general spatial neighbour relationships not already covered as strand adjacency or helix packing. In parallel we have developed an entity relationship data model for TOPS diagram data, and implemented this as a MySQL (http://www.mysql.com/) relational database to replace the original flat file versions. The relational database also includes further annotation of the TOPS diagrams with information relating to protein sequence, structure and function.

The database and associated tools can be found at http://www.tops.leeds.ac.uk/. This website provides access to the database through several query interfaces suitable for different problems and users of differing degrees of experience. The main advantage of TOPS is the compact and highly simplified nature of the fold description, which enables better visualization of folding topology, very fast whole database searches and machine learning applications including motif extraction and multiple alignment. The database thus serves two main user constituencies: structural molecular biologists interested in visualization and database searching, and computer scientists or bioinformaticians interested in automated extraction of patterns relating protein sequence to folding topology and protein function.

*To whom correspondence should be addressed. Tel: +44 113 343 3116; Fax: +44 113 343 3167; Email: d.r.westhead@leeds.ac.uk

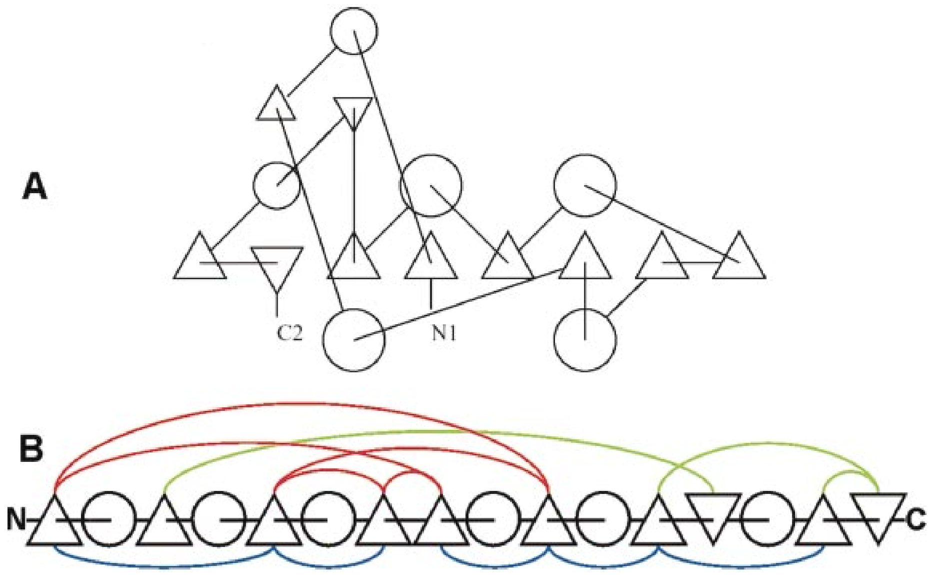

Figure 1. Topological representations of protein structures consider a sequence of SSEs, i.e. helices (circles) or strands (triangles), together with relationships like spatial adjacency within the fold and approximate orientation, neglecting details like the lengths of SSEs and loops. (**A**) A 2D TOPS cartoon for of 1ra9 (dihydrofolate reductase). TOPS cartoons are pseudo-2D schematic abstractions, where the third dimension is implied, since SSEs are considered to have an approximate direction of 'up' or 'down' (connecting lines drawn to the centre of the symbol indicate connection to the top, and those drawn to the edge indicate connection to the base). Direction information for strands is duplicated, upward pointing triangles indicating 'up' strands and vice versa. Adjacent strand pairs are connected by H-bonds, being parallel or anti-parallel. Chiralities between parallel strands are also implicit. (**B**) A TOPS diagram of 1ra9. Hydrogen bonds and supersecondary chiralities are shown explicitly (parallel in red, anti-parallel in green, right-handed chiralities in blue).

DATABASE CONTENT

The detailed data model for the TOPS database showing all tables and attributes is given in the summary schema shown in figure S1 in Supplementary Material. In total the database contains 20 tables. The information in the database can be classified into three main categories: basic information about the protein, topological description and visualization of the protein fold, and enhanced information about protein function. Basic information about the protein includes any available naming information, experimental details about structure determination and the amino acid sequence. This information is extracted directly from the PDB. Information in the other categories is described in more detail below.

The topological information in the database is summarized in the TOPS diagram shown in Figure 1. It comprises a set of SSEs connected in sequence from amino- (N) to carboxy- (C) terminus. In addition to the sequence, four sets of relationships (hydrogen bonding, helix packing, chirality and neighbour relationships) relate secondary structure elements. The TOPS diagram is thus a mathematical graph with the SSEs as nodes and five edge sets including the directed N- to C-terminal sequential edges, and the four other undirected pairwise relationships between SSEs.

The β strands that are adjacent in the folded protein are connected by hydrogen bond relationships. These can be parallel or anti-parallel according to the pattern of underlying atomic hydrogen bonds connecting the two strands in question. By analogy, pairs of α helices that are in atomic contact (when atoms of at least three residues of each helix form atomic contacts between the two helices) are connected by helix packing relationships. Helix packing data were produced by Helix Packing Pair (5), which defines helical axes and describes helical geometry as straight, curved or kinked (6), and subsequently calculates the helix packing angle as the angle between the skew lines representing helical axes. Two packing angles are defined: one calculated from the helical axes local to the point of contact, the other based on a global best fit straight line axis for the complete helix. These only differ significantly in the cases of highly curved or kinked helices. SSE contacts, other than hydrogen bonds and helix packing, are represented as general neighbour relationships in the database.

Supersecondary structure chirality is also represented as a pairwise relationship between SSEs. For instance, in the case of βαβ units the chirality is represented as a relationship, labelled as right- or left-handed, between the two parallel strands. It is well known that this connection between β strands, which might contain one or more α helices or other SSEs, is strongly preferred to be right-handed in natural protein structures (7). Other similar chirality relationships are contained in the database, e.g. the chiralities of connections between parallel helices.

As well as basic topological information the database contains the 2D layout of the TOPS cartoon visualization aids. This is calculated automatically by the TOPS program. The optimal layout problem for TOPS cartoons is difficult (1) and we estimate that it is successful in ~80% of cases. The layout problem is made easier if the structure is first divided into structural domains, and the database schema is sufficiently flexible to hold several possible divisions of each structure into domains. Currently the domain definitions from CATH (8) and SCOP (9) databases are present. In addition to domain definitions, the database also holds the CATH and SCOP classification hierarchy. The data of a manually inspected Atlas (3) of selected domains is also included in the database.

The database is constantly updated by the creation of TOPS Chain data for all newly submitted protein structures that are not currently included in CATH/SCOP or the Atlas releases. TOPS cartoons of all these sources can be viewed and/or edited, using TOPS website facilities.

At the time of writing optimal cartoon layouts are not available for every structure in the database, but in cases where the layout is unsatisfactory the website also provides a cartoon server, where the user can submit their own structure, or an existing PDB structure, for cartoon calculation. This site also contains a TOPS diagram editor, implemented as a Java Applet, to enable the user to improve the aesthetic appearance of the cartoons.

The PDB is the main source of data from the database, which at the present time (August 2003) contains ~22 500 structures. New structures are added to the database by an analysis pipeline, beginning with secondary structure definition by the Dictionary of Secondary Structure of Proteins (DSSP) (10), the TOPS program and several other programs that calculate or extract the information described above.

QUERYING THE DATABASE

Queries of the database can be divided into the categories listed below, each implemented as a separate web interface:

(i) simple queries;
(ii) advanced queries using TOPS topological patterns;
(iii) structural comparison and similarity searches;
(iv) viewing TOPS cartoons for visualization.

The simple query interface, which is the best starting point for a novice user, comprises a number of ‘canned’ queries. These can be characterized as queries that can be implemented in a straightforward way using the SQL relational database query language. ‘Canned’ queries cover authors, PDB ID, chain ID, domain ID, source of classification, βαβ units of defined chirality, pairs of helices packing in the angular range $X<\theta<Y$ and SSEs binding ligands (peptides, nucleic acids or other compounds). Typically the queries allow the selection of data, subject to user-defined constraints, from user-defined subsets of the database. For example, a query to extract βαβ units of defined chirality, limited to the homologous superfamily representatives from the CATH database, revealed that 1.6% of βαβ are left-handed. This interface also allows queries linking topological and functional information.

The simple query interface allows simple topological queries such as the extraction of β hairpins, or βαβ units, that can also be visualized as TOPS cartoons. However, when topological queries become more complex the SQL involved in formulating the query becomes increasingly unwieldy. Such queries are better implemented using customized graph matching methods (2), and require the advanced interface using topological patterns. These patterns are described in detail elsewhere (2), but, for example, the pattern below defines a β 3-meander:

$$V = \beta_{+1}\text{-}(0,0)\text{-}\beta_{-2}\ \text{-}(0,0)\text{-}\beta_{+3}$$
$$H = \{(\beta_{+1},A,\beta_{-2}\ ),(\beta_{-2},A,\beta_{+3})\}$$
$$C = \varnothing$$

Here, a V pattern defines a sequence of three β strands, with connections containing no other SSEs, the second being antiparallel (–) to the others (+). The H pattern indicates antiparallel (A) hydrogen bond relationships between strands 1 and 2, and 2 and 3. The C pattern is empty, indicating no chirality relationships. In general, the patterns define a sequence of SSEs along with their topological relationships and the numbers of allowed intervening (‘inserted’) SSEs. Using such patterns it is possible to define topological queries of high complexity.

Structure comparison and database similarity search at the topological level have been implemented using a novel machine learning methodology (11). The user interface allows the user to compare two structures from the database or submitted as PDB files, or to carry out a similarity search of the database for a single input query structure.

Finally, the user interested simply in seeing TOPS cartoons to visualize protein folding topologies can access these using the fourth interface to the database. Also on the site is the original Atlas of high-quality TOPS cartoons (3) which we produced by manually checking, and editing if necessary, cartoons for a representative set of PDB structures. The old Atlas is still available on the site but it is unlikely to be updated. Our policy is to provide the best possible automatically generated cartoons for all PDB structures, and users who wish for perfect cartoons of published PDB structures or their own unpublished data, will be able to produce and edit them as necessary using the cartoon editor (see above). The new database may also be queried over the web using a standard URL which contains a PDB ID: http://www.tops.leeds.ac.uk:8080/tops/find/pdb_id.html or http://www.tops.leeds.ac.uk:8080/tops/find/1nfk.html. Any other database wishing to link to ours may do so using the above URL.

Automatically generated TOPS chain cartoons are exceptionally useful for automated domain definition, as they group SSEs of the same domain. These groups can be used as the basis for domain definitions. Users can generate data for domains, based on these or their own domain definitions.

FUTURE DIRECTIONS

The list of queries is constantly growing, according to the needs of the TOPS users. Topological descriptions will be annotated with function information of SSEs, including EC numbers and atomic contacts with all types of bound ligand, as cofactors, metal ions, small organic molecules, oligopeptides (<20 residues) and oligonucleotides. SSE–ligand interactions will be categorized as covalent, electrostatic, hydrogen and van der Waals, or in the case of SSEs bound to nucleic acids as sequence specific/non-specific or RNA/DNA specific.

We will construct a TOPS Java-3D viewer to facilitate the display of the SSEs of the protein domains, and manipulating TOPS flat (2D) diagrams in three dimensions—allowing rotation and scaling. This will be extended to a full 3D viewer, which will effectively reconstruct a highly idealized form of the original structure. This will have the advantage of simplicity of representation, including information that cannot be visualized in two dimensions, e.g. the lengths of the SSEs will be proportional to their actual length.

SUPPLEMENTARY MATERIAL

An entity relation diagram of the TOPS database is available as Supplementary Material at NAR Online.

ACKNOWLEDGEMENTS

I.M. and G.M.T. were supported by the BBSRC/EPSRC.

REFERENCES

1. Westhead,D.R., Slidel,T.W., Flores,T.P. and Thornton,J.M. (1999) Protein structural topology: Automated analysis and diagrammatic representation. *Protein Sci.*, **8**, 897–904.
2. Gilbert,D., Westhead,D., Nagano,N. and Thornton,J. (1999) Motif-based searching in TOPS protein topology databases. *Bioinformatics*, **15**, 317–326.
3. Westhead,D.R., Hatton,D.C. and Thornton,J.M. (1998) An atlas of protein topology cartoons available on the World-Wide Web. *Trends Biochem. Sci.*, **23**, 35–36.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
5. Dalton,J.A.R., Michalopoulos,I. and Westhead,D.R. (2003) Calculation of helix packing angles in protein structures. *Bioinformatics*, **19**, 1298–1299.
6. Bansal,M., Kumar,S. and Velavan,R. (2000) HELANAL: a program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn.*, **17**, 811–819.
7. Sternberg,M.J. and Thornton,J.M. (1976) On the conformation of proteins: the handedness of the β-strand–α-helix–β-strand unit. *J. Mol. Biol.*, **105**, 367–382.
8. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
9. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
10. Kabsch,W. and Sander,C. (1983) Dictionary of Protein Secondary Structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
11. Gilbert,D., Westhead,D., Viksna,J. and Thornton,J. (2001) A computer system to perform structure comparison using TOPS representations of protein structure. *Comput. Chem.*, **26**, 23–30.